INTERNET DOCUMENT INFORMATION FORM

A . Report Title: Final report COOP 3D ARPA Experiment 109 National Center for Atmospheric Research

B. DATE Report Downloaded From the Internet 9/22/98

C. Report's Point of Contact: (Name, Organization, Address, Office Symbol, & Ph #):

Nasa Lewis Research Center
21000 Brookpark Road
Cleveland, OH 44135-3127
ATTN: Doug Hoder (216) 433-8705

D. Currently Applicable Classification Level: Unclassified

E. Distribution Statement A: Approved for Public Release

F. The foregoing information was compiled and provided by:
DTIC-OCA, Initials: VM___ **Preparation Date**:_9/23/98________

The foregoing information should exactly correspond to the Title, Report Number, and the Date on the accompanying report document. If there are mismatches, or other questions, contact the above OCA Representative for resolution.

Final Report
COOP 3D ARPA Experiment 109
National Center for Atmospheric Research

EXERIMENTAL OBJECTIVE - NETWORK PERSPECTIVE:

To advance the knowledge and understanding of next generation high data rate telecommunications between widely distributed high performance computing systems via terrestrial fiber/satellite hybrid testbed network architecture.

ABSTRACT:

Coupled atmospheric and hydrodynamic forecast models were executed on the supercomputing resources of the National Center for Atmospheric Research (NCAR) in Boulder, Colorado and the Ohio Supercomputing Center (OSC)in Columbus, Ohio. respectively. The interoperation of the forecast models on these geographically diverse, high performance Cray platforms required the transfer of large three dimensional data sets at very high information rates. High capacity, terrestrial fiber optic transmission system technologies were integrated with those of an experimental high speed communications satellite in Geosynchronous Earth Orbit (GEO) to test the integration of the two systems. Operation over a spacecraft in GEO orbit required modification of the standard configuration of legacy data communications protocols to facilitate their ability to perform efficiently in the changing environment characteristic of a hybrid network. The success of this performance tuning enabled the use of such an architecture to facilitate high data rate, fiber optic quality data communications between high performance systems not accessible to standard terrestrial fiber transmission systems. Thus obviating the performance degradation often found in contemporary earth/satellite hybrids.

Introduction

Interworking of legacy supercomputer systems with the revitalized technologies embodied by the high power NASA Advanced Communications Technology Satellite (ACTS) required the integration of high performance terrestrial systems which were complementary with the spacecraft's expanded capabilities. Dissimilar physical layer technologies required integration, providing the opportunity to investigate their interoperability over high speed terrestrial and satellite media. Native High Performance Parallel Interface (HiPPI) traffic from the Cray supercomputers was converted to Synchronous Optical Network (SONET) framing. Asynchronous Transfer Mode (ATM) cells generated by the high performance workstations controlling model interaction and visualizing output were converted to HiPPI, then SONET. SONET was the physical layer for the ACTS satellite transmisson system, a key component in the hybrid architecture was the ability to facilitate the interoperation between physical layer technologies.

Another critical factor in the success of the hybrid model's interoperability was the optimization of Transmission Control Protocol/Internet Protocol (TCP/IP) for the high performance satellite channel. The high data rates afforded by SONET combined with the great latency (delay) of the space segment exceeded the domain of classical TCP functionality. This is manifested in the delay for the acknowledgement of a transmitted packet being great enough for the source's transmission window to timeout well before source packets reach the destination. The net result is a very inefficient "stop and wait" state on the source, effectively shutting off the flow of data between the application, the kernel and the transmission media. The link remains idle until an acknowledgement is received by the source which then transmits another packet. Performance is impacted as throughput declines radically from the source's inability to keep the channel full; the bit length of transmitted data being far less than that which the channel can accommodate. Performance enhancements to the TCP Automatic Retransmission Request (ARQ) or sliding window were taken advantage of to match the performance of TCP to the hybrid network.

IP->HiPPI->SONET Physical Layer Configuration

In local environments where the time-distance separation between machines is slight, HiPPI interfaces and HiPPI switches may directly perform the interconnection of end systems. Wide area terrestrial interoperability is then facilitated by layered protocol suites suited for longer distances such as TCP/IP which is mapped to the HiPPI stream providing either connectionless datagram service (UDP) or connection-oriented reliability (TCP).

Model data sets transferred by the Cray's HiPPI interaces, via a HiPPI switch were converted to a serial SONET stream by a HiPPI/SONET gateway developed at the Los Alamos National Laboratory (LANL). This stream was framed in SONET OC-3c (concatenated pointer, 155.52 mbps) Synchronous Payload Envelopes (SPE) and transferred to the ACTS ground segment High Data Rate (HDR) terminal via Single Mode optical fiber. SONET Section and Line Overhead was terminated by the HDR prior to transmission and regenerated by the receiving HDR's SONET section.

IP->ATM->HiPPI->SONET Physical Layer Configuration

The modeling components were linked using Parallel Virtual Machine (PVM). PVM is a software library which provides a uniform, parallel computing architecture which is independent of the underlying hardware and network topology. PVM also linked the models to the data managers, software written to filter and route the model output streams. This data was fed to visualization components which were executed on high performance Silicon Graphics Inc. (SGI) workstations. Control and evaluation of the simulation was afforded by this component. Collaboration among the scientific researchers was facilitated by packet-based (IP) over ATM video conferencing at each site.

The video conferencing and collaborative workstations were directly connected via ATM to Fore ASX-200 ATM switches. IP packets segmented into 53 byte cells by the workstation's ATM Network Interface Cards (NIC) were converted to HiPPI by a NetStar GigaRouter and routed through the HiPPI switch to the HiPPI/SONET gateway. As with the Cray output, this HiPPI stream was framed in SONET OC-3c SPE and transferred to the HDR terminal over Single Mode fiber. Section and Line Overhead being terminated by the transmitting HDR and regenerated by the receiver. SONET frames from the workstations were transmitted along with model data via the single OC-3c ACTS channel, completing the terrestrial complement of the hybrid.

TCP Performance

The ability of an end system to transmit data is ultimately limited by the information capacity of the transmission medium. Efficient use of the medium is achieved by maintaining transmission rates at or close to the maximum. The combination of this data rate capability and the round trip time (RTT) between source and destination specifies how much data is flowing at any instant between the sender and receiver.

TCP is a reliable, end to end connection oriented transport layer protocol which uses a sliding window based flow control system or ARQ to recover from loss or corruption of data over the medium. To achieve this, TCP requires the source to hold the data transmitted in buffer for a minimum of the time required to send the data to the destination and receive an acknowledgement from the receiver or the RTT. Should data be corrupted or lost the entire contents of the source buffer is resent.

Maximum performance is obtained from TCP not just from high information rates but from the product of the information rate and the RTT. This "bandwidth-delay" product is equivalent to the amount of unacknowledged data outstanding at any instant on the transmission medium. The bandwidth-delay product then corresponds to the minimum buffer size or window size which will keep the "pipe" or link full and provide adequate recovery to congestion or loss. The larger the window, the more data can be outstanding and the capacity of the data link maintained at or near maximum capacity.

TCP window size corresponds to the size of the socket buffer space or send and receive buffers in both the source and destination UNIX operating system kernels. During connection establishment the source and destination negotiate the size of this window, facilitating the smooth, continuous flow of data for the duration of the connection. To provide efficient use of high capacity links with high latency, very large window sizes are required.

In the original TCP specification, RFC 793, the TCP header contains a 16 bit window size field corresponding to the receiver's window size. The 16 bit field can support a window size of 2E16, a maximum of 64 KBytes. RFC 1323 prescribes a window scalability option for the TCP header which can accommodate larger window sizes, up to 1 Gbyte. This option can improve the performance of modern networks with high bandwidth-delay products. The extension maps the standard 16 bit window size field to a 32 bit value and uses the window scale option to bit-shift this value, producing a new maximum window size value.

The window scale option occupies 3 bytes in the TCP header, it specifies the type of option as window scale and the second 3 bytes the length of the option and the shift count. The window scale indicates the sender is able to accept send and receive buffer or window scaling and sends the scale factor to the receiver. The window scale is a log base2 value and the shift count is the number of bits the receiver's window value is to be right shifted. Right shift applies to the default TCP window specified in the TCP header. Values less than the 2E16 maximum will only be right shifted by the shift factor.

An application may set a larger window size with the setsockopt call, based on the available buffer space of the operating system kernel. The implementation of window scale will then determine the appropriate shift factor. The maximum window shift could be obtained starting with a default maximum window size of 2E16 and a scale factor of 14. This results in the maximum window size of 1 GByte (2E16 * 2E14 = 2E30 = 1.073 Gbyte).

Application of this TCP performance extension requires that the operating system kernel of the source and destination include the extensions to TCP performance detailed in RFC 1323. The maximum amount of socket buffer space available in the operating system kernel must be great enough to accommodate the window scale factor anticipated. The RTT of the data link and the maximum information rate must be known to facilitate performance enhancement using the window scale option to adjust window size.

System Integration and Test Configurations

Due to the complexity of the architecture, various levels of system integration were performed. Commensurate continuity and performance tests were made to validate progress and functionality of the physical layer integration and the TCP performance enhancements prior to advancement to the next level. The levels were:

1. Earth Station Installation
2. Network Hardware Installation of GigaRouter and HiPPI/SONET gateway
3. Window Size Optimization
4. Satellite Loopback Tests
5. End-to-end connectivity over ACTS

As all physical layer data streams were converted to SONET for transmission over the ACTS spacecraft, interoperation of the two HiPPI/SONET gateways was verified at NCAR prior to the shipment and installation of the second gateway at the OSC site. Various configurations of both HiPPI/SONET gateways were tested in loopback with single Cray connectivity to test continuity and validate raw HiPPI and HiPPI/SONET performance for each. The gateways were then tested between two local Crays to simulate end-to-end connectivity involving both gateways and different machines. Finally each gateway was tested between two local Crays via a local loopback at the NCAR HDR. HiPPI performance was validated by a simple program which writes 10 Mbyte buffers of raw HiPPI data across logical interfaces on NSC PS-32 HiPPI switch to a single or pair of Crays.

HiPPI tests were made for the following HiPPI/SONET configurations.

1. Single Cray looped through each gateway individually via HiPPI switch.
2. Single Cray looped through both gateways back-to-back via HiPPI switch.
3. Cray to Cray via both HiPPI/SONET gateways back-to back via HiPPI switch.
4. Cray to Cray via HiPPI switch, single HiPPI/SONET gateway and HDR digital terminal in loopback.

Terrestrial and spacecraft TCP performance baselines were established prior to TCP window optimization. Tests were conducted between two local Crays at NCAR interconnected similarly to HiPPI test configurations and to the spacecraft in loopback (bent pipe).

TCP performance tests were made for the following HiPPI/SONET and spacecraft configurations.

1. Cray to Cray via single HiPPI/SONET gateway in loopback via HiPPI switch.
2. Cray to Cray via both HiPPI/SONET gateways back-to back via HiPPI switch.
3. Cray to Cray via HiPPI switch, single HiPPI/SONET gateway and HDR digital terminal in loopback.
4. Cray to Cray via HiPPI switch, single HiPPI/SONET gateway over the ACTS OC-3c bent pipe.

Once validated, latency for the round trip spacecraft OC-3 channel was incorporated into a Long Link Emulator (LLE) to simulate the satellite delay for performance enhancement and application development during periods when spacecraft time was not available.

1. Cray to Cray via both HiPPI/SONET gateways and Long Link Emulator via HiPPI switch.
2. Cray to Cray via single HiPPI/SONET gateway, LLE in loopback via HiPPI switch.

Full end-to-end connectivity between NCAR and OSC was established after completion of all above integration tests. Successful NCAR Cray to OSC Cray connectivity was made through identical configurations at each site.

TCP Performance Tuning

The required TCP window size for the Crays and hence the window shift was determined from the bandwidth-delay product of the hybrid. The round trip time was calculated empirically then validated.

A round trip for a packet and its corresponding acknowledgement requires two satellite hops. The packet traverses the link once enroute to the destination where if received correctly, elicits an acknowledgement. The acknowledgement then traverses the link back to the source, completing the round trip. The calculation of the time required for this round trip between two Cray machines at NCAR interconnected via the satellite loopback was made based on the the one-way propagation time between the NCAR HDR and the spacecraft. The documented location of the NCAR HDR terminal was used in this calculation.

39 degrees 58 minutes 39 seconds north latitude
105 degrees 16 minutes 28 seconds west longitude
6113 feet above mean sea level

For the purpose of calculation latitude was rounded to 40 degrees, the radius of the earth was taken to be 3960 miles and the orbital altitude of the spacecraft was assumed to be 22300 miles above the earth's equator. The speed of light, c was taken to be 186400 miles/second in the vacuum of space and the atmosphere. The law of sines was used given a triangle formed by the earth segment antenna (point B), the center of the earth (point A) and the spacecraft (point C). The sides of the triangle opposite these angles are a,b and c respectively.

$$a/\sin 40 = 3960/\sin b = 22300 + 3960/\sin c$$

$$a + b + c = 180, \text{ if } a = 40, \text{ then,}$$
$$40 + b + c = 180$$
$$b + c = 180 - 40$$
$$b + c = 140$$
$$c = 140 - b$$

3960/sinb = 26260/sin(140 - b)
b = 6.25

a/sin40 = 3960/sin6.25
a = 23365 miles

Delay = 23365 miles/186400 miles/sec
= .125349 sec to spacecraft one way
= 2 (.125349 sec)
= .250697 sec

Round trip time = 2(.250697 sec)
= .501394 sec
= 501 ms

Validation of the empirical RTT was made using a series of Internet Control Message Protocol (ICMP) queries (ping) sent between a Cray Y-MP8 and a Cray EL-92 over the spacecraft bent pipe.

```
16-aztec% /etc/ping echo-h
PING echo-h.ucar.edu: 56 data bytes
64 bytes from 128.117.5.5: icmp_seq=0. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=1. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=2. time=541. ms
64 bytes from 128.117.5.5: icmp_seq=3. time=544. ms
64 bytes from 128.117.5.5: icmp_seq=4. time=543. ms
64 bytes from 128.117.5.5: icmp_seq=5. time=542. ms
64 bytes from 128.117.5.5: icmp_seq=6. time=540. ms
64 bytes from 128.117.5.5: icmp_seq=7. time=540. ms
64 bytes from 128.117.5.5: icmp_seq=8. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=9. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=10. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=11. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=12. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=13. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=14. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=15. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=16. time=540. ms
64 bytes from 128.117.5.5: icmp_seq=17. time=541. ms
64 bytes from 128.117.5.5: icmp_seq=18. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=19. time=539. ms
64 bytes from 128.117.5.5: icmp_seq=20. time=539. ms

----echo-h.ucar.edu PING Statistics----
21 packets transmitted, 21 packets received, 0% packet loss
```

```
round-trip (ms)  min/avg/max = 539/539/544
```

Local pings between a Cray Y-MP8 and a Cray EL-92 over the HiPPI/SONET gateway in loopback.

```
63-echo% ping aztec2-h.acts
PING aztec2-h.acts.ucar.edu: 56 data bytes
64 bytes from 192.157.4.32: icmp_seq=0. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=1. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=2. time=6. ms
64 bytes from 192.157.4.32: icmp_seq=3. time=5. ms
64 bytes from 192.157.4.32: icmp_seq=4. time=3. ms
64 bytes from 192.157.4.32: icmp_seq=5. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=6. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=7. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=8. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=9. time=7. ms
64 bytes from 192.157.4.32: icmp_seq=10. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=11. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=12. time=5. ms
64 bytes from 192.157.4.32: icmp_seq=13. time=3. ms
64 bytes from 192.157.4.32: icmp_seq=14. time=4. ms
64 bytes from 192.157.4.32: icmp_seq=15. time=4. ms

----aztec2-h.acts.ucar.edu PING Statistics----
16 packets transmitted, 16 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 3/4/7
```

The average RTT of 539 ms was compared to the calculated value of 501 ms and the assumption was made that the additional 38 ms was the delay imposed somewhere between the Crays and the NCAR HDR terminal. One-way delays observed in raw HiPPI tests between the Crays with the HDR digital terminal in loopback revealed fairly consistent first packet delays of 20 ms. Local pings across the HiPPI/SONET gateway in loopback and the LLE set for 0 ms emulation delay, revealed very small delays suggesting the bulk of the additional 39 ms delays was the result of the involvement of two trips through the HDR terminals; approximately 20 ms for the transmitted packet in one direction and another 20 ms for the acknowledgement in the opposite direction.

The data below shows the delay for the first raw HiPPI packet to return from the HDR loopback consistently in the 20 ms range.

```
Output channel -- HXCF_HIPPI is set
HXCF_HDR is set (user buffer has FP header)
HXCF_IND is zero (I-field not in user buffer)
HXCF_ISB is zero (Short burst at end)
```

```
 Data length = 32768 bytes.
 FP header: 8580001800007fe0
 I-field is 702a3d2
 HXC_SET for output OK.
 Write of 10485760 bytes completed successfully
 Elapsed microsecs = 935094
 Fastest single block was 112.267 Mbits/sec
 Slowest single block was 46.463 Mbits/sec
 Overall data rate was   89.709 Mbits/sec
Catch completed successfully
 ..First packet delay was 20876 microsec.
 ..Overall data rate with delay was   87.733 Mbits/sec
 ..Overall data rate without delay was   89.691 Mbits/sec

Output channel -- HXCF_HIPPI is set
HXCF_HDR is set (user buffer has FP header)
HXCF_IND is zero (I-field not in user buffer)
HXCF_ISB is zero (Short burst at end)
 Data length = 32768 bytes.
 FP header: 8580001800007fe0
 I-field is 702a3d1
 HXC_SET for output OK.
 Write of 10485760 bytes completed successfully
 Elapsed microsecs = 882882
 Fastest single block was 112.993 Mbits/sec
 Slowest single block was 29.008 Mbits/sec
 Overall data rate was   95.014 Mbits/sec
Catch completed successfully
 ..First packet delay was 19809 microsec.
 ..Overall data rate with delay was   91.847 Mbits/sec
 ..Overall data rate without delay was   93.883 Mbits/sec

Output channel -- HXCF_HIPPI is set
HXCF_HDR is set (user buffer has FP header)
HXCF_IND is zero (I-field not in user buffer)
HXCF_ISB is zero (Short burst at end)
 Data length = 65536 bytes.
 FP header: 858000180000ffe0
 I-field is 702a3d1
 HXC_SET for output OK.
 Write of 10485760 bytes completed successfully
 Elapsed microsecs = 749291
 Fastest single block was 122.269 Mbits/sec
 Slowest single block was 40.016 Mbits/sec
 Overall data rate was   111.954 Mbits/sec
Catch completed successfully
```

```
..First packet delay was 21504 microsec.
..Overall data rate with delay was    109.334 Mbits/sec
..Overall data rate without delay was    112.486 Mbits/sec
```

The SONET OC-3 information rate is 155.52 mbps, however Section and Line Overhead are terminated by the HDR but the Path Overhead is still present in the SPE. The nettest/nettestd performance measurements are made at the application level, so a more conservative rate of 135 mbps was used in the bandwidth-delay product calculation to account for this overhead.

(135E6 b/s) * (539 ms) = 72765000 bits/8 bits/byte = 9095625 bytes

The bandwidth-delay product above specifies the minimum send/receive buffer or window size required for optimum TCP performance on the hybrid network. This value specifies a window shift of 8 (2E8). While a window size of 9095625 bytes is optimal, the window shift will be set for the next larger increment. A shift of 8 would yield a much larger buffer that the computed value, while a shift of only 7 would not provide one that is large enough. Specification of the correct buffer size in the nettest utility with the -b option will set the correct buffer size. A window shift of 8 will accommodate that size and any size up the maximum shift value fo (2E16) * (2E8).

A modified version of the Cray UNICOS TCP test utility nettest and nettestd was used to measure the effect of the window size and window size changes on throughput performance. The nettest/nettestd utility performs client and server functions for measuring network throughput of interprocess communication. The nettest program establishes a connection with the nettestd program which performs the server function, waiting for the nettest client to initiate the process communications. As with any TCP connection, the window scale option is sent at connection establishment in the <SYN> segment. Thus the window scale value is fixed when the connection is made and remains for the duration of the connection. The nettest program writes a number of bytes to the nettestd program which reads a number of bytes and reports the throughput. Nettestd in turn writes a number of bytes to nettest which reads them and also reports the performance. The preformance of the two processes is averaged to disclose an average data throughput rate for the connection.

Modification to the nettestd source code was required because Cray's nettestd does not have the capability to allow the user to set the the window size or shift factor. Rather the default maximum window size defined in the operating system kernel (in this case 2E16 = 65536 byte) is used by nettestd. To overcome this limitation it was put forth that perhaps during connection establishment nettestd "spoofs" the server that the size of its receive buffer is the same as the sender's, a requirement for smooth data flow in both directions. The connection is established but an imbalance in buffer sizes between the source and destination may results in asymmetric transfer rates observed between the data sent between the client and the server.

While unable to verify that this spoofing is in fact occurring, modification of the nettestd code allows the user to define the window shift for both nettest and nettestd by defining the -s and -b options in the nettest program. Upon connection establishment the specified window sizes are set on both client and server, resulting in symmetric data transfers between client and server assuming similar loads on the two machines.

Results

Performance tests using the modified nettest/nettestd programs were executed for the various configurations. The results of the tests between different Cray platforms were validated against each other for bent pipe tests as well as full connectivity end-to-end tests between NCAR and OSC. Tests were made using various combinations of window shift and buffer size to determine and validate optimal TCP performance parameters.

Data below describe tests between a Cray EL-92 and a Cray J-916 at NCAR. The window scale was set for a shift of 8 and the send/receive buffer size is set for 10000000 in the setsockopt call. Large amounts of data are required to keep the "pipe full" and ensure link efficiency. The number of read or write operations and the number bytes to be read or written with each operation can be varied to affect the amount of data.

In the set of data below the number of times data was read/written was set to 100 and the number of bytes for each of the operations was 3000000. A disparity in performance is evident between the results of these tests and tests immediately following where the number of times data was read/written was set to 1000. It is not clear why more optimal performance is observed when the amount of data is increased by an order of magnitude. The amount of data transmitted is sufficient to "fill the pipe" in either case. While this is puzzling the total 'Real' time to effect the transfer increases by a only factor of seven, indicating more efficient use of the channel.

aztec=client
echo=server

```
64-aztec% rtx nettest -s 8 -b 10000000 echo-h.acts.ucar.edu 100 3000000

Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 100*3000000 bytes from     aztec to echo-h.acts.ucar.edu
         Real  System            User         Kbyte   Mbit(K2) mbit(1+E6)
  write 28.9430  4.9760 (17.2%)  0.0013 ( 0.0%) 10122.25  79.080    82.921
  read 30.5650  4.3535 (14.2%)  0.0626 ( 0.2%) 9585.11  74.884    78.521
   r/w 59.5080  9.3295 (15.7%)  0.0639 ( 0.1%) 9846.36  76.925    80.661
```

echo=client
aztec=server

```
83-echo% rtx nettest -s 8 -b 10000000 aztec2-h.acts.ucar.edu 100 3000000

Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 100*3000000 bytes from      echo to aztec2-h.acts.ucar.edu
        Real  System            User         Kbyte   Mbit(K2) mbit(1+E6)
 write 29.5932  9.1337 (30.9%)  0.0020 ( 0.0%) 9899.87  77.343    81.100
  read 30.3065  6.8523 (22.6%)  0.0977 ( 0.3%) 9666.85  75.522    79.191
   r/w 59.8997 15.9860 (26.7%)  0.0997 ( 0.2%) 9781.97  76.422    80.134
```

aztec=client
echo=server

```
72-aztec% rtx nettest -s 8 -b 10000000 echo-h.acts.ucar.edu 1000 3000000

Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 1000*3000000 bytes from     aztec to echo-h.acts.ucar.edu
        Real  System            User         Kbyte   Mbit(K2) mbit(1+E6)
 write 197.3138 52.9243 (26.8%)  0.0124 ( 0.0%) 14847.86 115.999   121.634
  read 215.8563 43.6969 (20.2%)  0.6384 ( 0.3%) 13572.40 106.034   111.185
   r/w 413.1701 96.6213 (23.4%)  0.6507 ( 0.2%) 14181.51 110.793   116.175
```

echo=client
aztec=server

```
91-echo% rtx nettest -s 8 -b 10000000 aztec2-h.acts.ucar.edu 1000 3000000

Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 1000*3000000 bytes from      echo to aztec2-h.acts.ucar.edu
        Real  System            User         Kbyte   Mbit(K2) mbit(1+E6)
 write 236.7876 95.9306 (40.5%)  0.0201 ( 0.0%) 12372.64  96.661   101.357
  read 196.6013 68.4667 (34.8%)  0.9884 ( 0.5%) 14901.67 116.419   122.075
   r/w 433.3889 164.3974 (37.9%)  1.0085 ( 0.2%) 13519.90 105.624   110.755
```

This phenomena was observed in both bent pipe configurations between machines at

NCAR and in later end-to-end tests between NCAR and OSC Crays.

aztec=client
osca=server

```
20-aztec% rtx nettest -s 8 -b 10000000 osca3.acts.ucar.edu 100 3000000
```

```
Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 100*3000000 bytes from     aztec to osca3.acts.ucar.edu
        Real  System           User         Kbyte  Mbit(K2) mbit(1+E6)
 write 28.8057  4.9513 (17.2%)  0.0013 ( 0.0%) 10170.51  79.457    83.317
  read 30.0736  4.4901 (14.9%)  0.0623 ( 0.2%) 9741.73  76.107    79.804
   r/w 58.8793  9.4414 (16.0%)  0.0635 ( 0.1%) 9951.51  77.746    81.523
```

ztec=client
osca=server

```
22-aztec% rtx nettest -s 8 -b 10000000 osca3.acts.ucar.edu 1000 3000000
-a temp
```

```
Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 1000*3000000 bytes from     aztec to osca3.acts.ucar.edu
        Real  System           User         Kbyte  Mbit(K2) mbit(1+E6)
 write 205.1117 53.1284 (25.9%)  0.0126 ( 0.0%) 14283.38 111.589   117.009
  read 194.9550 45.1127 (23.1%)  0.6294 ( 0.3%) 15027.50 117.402   123.105
   r/w 400.0667 98.2411 (24.6%)  0.6420 ( 0.2%) 14646.00 114.422   119.980
```

In spite of this anomaly, 1000 read/write operations appeared to be the minimum number of operations which would produce optimal throughput performance. It was used for all subsequent tests.

Batteries of nettest/nettestd suites were run to validate the empirical optimum of window shift 8 and a send/receive buffer size of 9095625 bytes. Data below supports the notion that window shifts and buffer sizes above or below theoretical optimum result in inferior performance.

(135E6 b/s) * (539 ms) = 72765000 b /8 b/B = 9095625 Bytes requires a window shift = 8

Window size is .5 optimal

```
311-aztec% rtx nettest -s 7 -b 4547813 echo-h 1000 1000000
```

```
Final SO_SNDBUF=4547813
Final SO_RCVBUF=4547813
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 7, recvwindshift= 7.
Transfer: 1000*1000000 bytes from     aztec to    echo-h
        Real  System         User        Kbyte   Mbit(K2) mbit(1+E6)
 write 124.9139 16.5351 (13.2%)  0.0098 ( 0.0%) 7817.89  61.077    64.044
  read 124.9731 10.9561 ( 8.8%)  0.1682 ( 0.1%) 7814.18  61.048    64.014
   r/w 249.8870 27.4912 (11.0%)  0.1781 ( 0.1%) 7816.03  61.063    64.029
```

Window size is optimal

```
293-aztec% rtx nettest -s 8 -b 9095625 echo-h 1000 1000000
```

```
Final SO_SNDBUF=9095625
Final SO_RCVBUF=9095625
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 1000*1000000 bytes from     aztec to    echo-h
        Real  System         User        Kbyte   Mbit(K2) mbit(1+E6)
 write 72.0108 18.6173 (25.9%)  0.0098 ( 0.0%) 13561.34 105.948   111.094
  read 76.0022 10.5164 (13.8%)  0.1684 ( 0.2%) 12849.13 100.384   105.260
   r/w 148.0130 29.1338 (19.7%)  0.1782 ( 0.1%) 13195.63 103.091   108.099
```

Window size 1.5 optimal

```
331  rtx nettest -s 8 -b 13643438 echo-h 1000 1000000
```

```
Final SO_SNDBUF=12000000
Final SO_RCVBUF=12000000
```

From: fair@niwot.scd.ucar.EDU (Chris Fair)
Subject: COOP 3D Final Report
To: dhoder@lerc.nasa.gov
Date: Fri, 12 Jan 1996 14:15:10 -0700 (MST)
MIME-Version: 1.0

```
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 1000*1000000 bytes from     aztec to    echo-h
        Real  System         User        Kbyte   Mbit(K2) mbit(1+E6)
 write 131.4753 15.5650 (11.8%)  0.0098 ( 0.0%) 7427.72  58.029    60.848
  read 138.7509 10.7622 ( 7.8%)  0.1711 ( 0.1%) 7038.24  54.986    57.657
   r/w 270.2263 26.3272 ( 9.7%)  0.1809 ( 0.1%) 7227.74  56.467    59.210
```

The results indicate that the optimal window shift and socket buffer size for the ACTS channel was a shift of 8 and a send/receive buffer size of no less than 9095625 bytes, validating the computed results. 10000000 bytes was chosen as a simple figure to work with and as the data below illustrates, was a valid assumption.

```
82-echo% rtx nettest -s 8 -b 10000000 aztec2-h 1000 3000000

Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 1000*3000000 bytes from      echo to  aztec2-h
        Real  System            User         Kbyte   Mbit(K̄2) mbit(1+E6)
 write 202.3199 97.4109 (48.1%)  0.0202 ( 0.0%) 14480.47 113.129   118.624
  read 195.3330 69.4343 (35.5%)  0.9929 ( 0.5%) 14998.42 117.175   122.867
   r/w 397.6530 166.8452 (42.0%)  1.0131 ( 0.3%) 14734.90 115.116   120.708
```

Repeated nettest/nettestd suites were executed using the validated window scale parameters in the bent pipe configuration. Consistent performance in the 120 mbps range was noted.

Having validated Cray to Cray performance in bent pipe configuration for the two machines at NCAR, validation tests using nettest/nettestd between the Cray J-916 at NCAR and the Cray Y-MP8 at OSC in end-to-end configuration were commenced. This would be the final test configuration and the one that the coupled atmospheric and hydrodynamic models would run on. The objective was to use all of the parameters from the bent pipe configurations tests to validate end-to-end connectivity performance.

```
OSCA$ nettest -s 8 -b 10000000 aztec2-h.acts.ucar.edu 1000 3000000

Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 1000*3000000 bytes from      OSCA to aztec2-h.acts.ucar.edu
        Real  System            User         Kbyte   Mbit(K̄2) mbit(1+E6)
 write 103.8431 11.1099 (10.7%)  0.0022 ( 0.0%) 14106.32 110.206   115.559
  read 104.6689  7.3552 ( 7.0%)  0.0896 ( 0.1%) 13995.03 109.336   114.647
   r/w 208.5119 18.4651 ( 8.9%)  0.0918 ( 0.0%) 14050.45 109.769   115.101
```

```
98-aztec% nettest -s 8 -b 10000000 osca3.acts.ucar.edu 1000 3000000

Final SO_SNDBUF=10000000
Final SO_RCVBUF=10000000
Final TR_SENDWNDSHIFT = 800000000000, sendwinshift = 8, recvwindshift= 8.
Transfer: 1000*3000000 bytes from     aztec to osca3.acts.ucar.edu
        Real  System            User         Kbyte   Mbit(K̄2) mbit(1+E6)
 write 103.8082 25.2396 (24.3%)  0.0063 ( 0.0%) 14111.06 110.243   115.598
  read 125.4580 22.2055 (17.7%)  0.3108 ( 0.2%) 11675.97  91.218    95.650
  r/w 229.2662 47.4451 (20.7%)  0.3171 ( 0.1%) 12778.54  99.832   104.682
```

As is evidenced above, TCP performance over the hybrid between the Cray J-916 at NCAR and the Cray Y-MP8 at OSC was as expected, validating the TCP Performance Extensions outlined in RFC 1323. These parameters were conveyed to the researchers for incorporation into the coupled model-PVM applications.

ATM->HiPPI Performance

Due to time constraints ATM to HiPPI performance over the hybrid was not validated. However with the exception of the Maximum Transmission Unit (MTU) or Maximum Segment Size (MMS) the performance parameters were the same as those between the Crays over the hybrid. Optimum performance is obtained in any environment by transmitting as large a packet (IP) as practicable. An MTU of 65 Kbytes for the HiPPI physical layer is used by the Crays while one of 9188 bytes is used by ATM for the video conferencing and collaborative workstations.

ATM to HiPPI performance was validated from the NCAR Cray J916 to the NCAR SGI Onyx. The path extended from the Cray, over HiPPI to the NetStar GigaRouter, to the SGI via ATM. The HiPPI stream from the Cray was converted to ATM and SONET by the NetStar GigaRouter. The success of this communications was the result of the use of dynamic MTU discovery on both machines. The correct MTU for the physical layer (9100 bytes) was automatically selected by both machines based on the installed configuration.

The following data examine the nettest/nettestd performance between the Cray J-916 and the SGI Onyx. The noticeably smaller window sizes reflect the low latency of the terrestrial fiber network at NCAR.

From the Cray J916 to the SGI Onyx

```
magic.47: nettest -b 37500 aztec2-h.acts 100 1000000
Transfer: 100*1000000 bytes from     magic to aztec2-h.acts

        Real  System          User        Kbyte  Mbit(K2) mbit(1+E6)
 write 15.9000  3.7700 (23.7%)  0.0100 ( 0.1%) 6141.90  47.984    50.314
  read 15.3600  5.2200 (34.0%)  0.1100 ( 0.7%) 6357.83  49.671    52.083
   r/w 31.2600  8.9900 (28.8%)  0.1200 ( 0.4%) 6248.00  48.813    51.184

magic.48: nettest -b 65535 aztec2-h.acts 100 1000000
Transfer: 100*1000000 bytes from     magic to aztec2-h.acts
        Real  System          User        Kbyte  Mbit(K2) mbit(1+E6)
 write 11.2000  3.8100 (34.0%)  0.0100 ( 0.1%) 8719.31  68.120    71.429
  read 12.7500  5.0100 (39.3%)  0.1100 ( 0.9%) 7659.31  59.838    62.745
   r/w 23.9500  8.8200 (36.8%)  0.1200 ( 0.5%) 8155.01  63.711    66.806
```

From the SGI Onyx to the Cray J-916

```
aztec.6: nettest -b 37500 magic.acts 100 1000000

Final SO_SNDBUF=37500
Final SO_RCVBUF=37500
Final TR_SENDWNDSHIFT = 0, sendwinshift = 0, recvwindshift= 0.
Transfer: 100*1000000 bytes from     aztec to magic.acts
         Real  System          User        Kbyte  Mbit(K2) mbit(1+E6)
  write 18.0001  6.5029 (36.1%)  0.0012 ( 0.0%) 5425.32  42.385    44.444
   read 15.7929  9.2299 (58.4%)  0.1281 ( 0.8%) 6183.55  48.309    50.656
    r/w 33.7930 15.7328 (46.6%)  0.1292 ( 0.4%) 5779.67  45.154    47.347
```

```
aztec.8: nettest -b 65535 magic.acts 100 1000000

Final SO_SNDBUF=65535
Final SO_RCVBUF=65535
Final TR_SENDWNDSHIFT = 0, sendwinshift = 0, recvwindshift= 0.
Transfer: 100*1000000 bytes from     aztec to magic.acts
         Real  System          User        Kbyte  Mbit(K2) mbit(1+E6)
  write 12.4164  6.2488 (50.3%)  0.0013 ( 0.0%) 7865.09  61.446    64.431
   read 13.3255  9.0434 (67.9%)  0.0996 ( 0.7%) 7328.50  57.254    60.035
    r/w 25.7420 15.2922 (59.4%)  0.1009 ( 0.4%) 7587.32  59.276    62.155
```

The performance in both directions, alternating client and server is consistent and shows marked performance increase with larger window sizes. Notice that the nettest/nettestd commands do not specify the -s option. This is due to the fact that the SGI will invoke the appropriate window shift when the -b option signals a socket buffer size greater that the default window size set in the kernel. The Cray however requires the -s option but not in this case as the window size does not exceed the default. Large socket buffers are not required due to a trivial bandwidth-delay product. For interactive control of the models via the hybrid by the SGI workstations however, the larger window sizes were required.

Problems and Recommendations

A. Difficulty integrating SONET components of the HiPPI/SONET gateway with the SONET section of the HDR digital terminals.

A major objective of the project was to investigate the interoperation of different physical layer technologies into the hybrid architecture. A prototype HiPPI/SONET gateway was integrated to serialize the parallel HiPPI datastreams for mapping to the SONET SPE of the earth station HDR's. A great deal of time was spent researching problems establishing loopback as well as end-to-end connectivity through this device. The result was a loss of scheduled time on the spacecraft due to earth segment unavailability. This had an overall adverse impact on the progress of the entire experiment. Project milestones were not reached, causing a slide to the right of the experimental time line.

It was proposed that this phenomenon was the result of synchronization problems between the gateway and the HDR. Numerous earth station components were replaced in an effort to locate this inconsistency. Toward the end of the project BBN, LANL, NCAR and OSC engineers determined that the cause was a continuous cyclic reset between the two components. A procedure was developed to circumvent what appeared to be a State Machine problem with the SONET section of the HDR terminals.

Upon spacecraft acquisition the HiPPI/SONET gateways must be reset in such a manner as to establish synchronization during a perceived time window of opportunity. Timing must be set to source on BOTH HiPPI/SONET gateways, each recovering timing from the other. The sequential order of synchronization was found to be critical. Once found, this procedure consistently established immediate end-to-end connectivity between the two sites. It was used with complete success for the duration of the experiment. Regrettably this discovery was made late in the life cycle of the experiment and could not recover time lost.

The procedure is outlined as follows:

1. Terminate all traffic over spacecraft link.
2. Turn off HiPPI/SONET gateway and FIFO at first site.
3. Leave HiPPI/SONET gateway and FIFO down at first site during power-off of the HiPPI/SONET gateway and FIFO at second site.
4. Restart FIFO at second site, and restart HiPPI/SONET gateway after FIFO.
5. Restart FIFO at first site, and restart HiPPI/SONET gateway after FIFO.
6. Do not attempt to transmit ANY data traffic until verification of HDR state and spacecraft (TDMA) acquisition at both sites.
7. If state of both HDR's not OK go back to step 2. and begin again.
8. If state of both HDR's is OK initiate data traffic and evaluate link performance.

B. Poor reliability and MTBF for critical earth segment hardware components.

Problems encountered with hardware reliability in both HDR earth stations resulted in the loss of time slots for critical end-to-end connectivity on the spacecraft. The OSC site experienced Traveling Wave Tube Amplifier (TWTA) failure twice during the June-October '95 period. Digital Terminal components at NCAR were frequently replaced to correct Network Processors, Timing and Control Card and SONET section failures. This component instability had an adverse impact on the schedule due to earth segment unavailability at critical experimental phases. In spite of this BBN was very responsive and effective in troubleshooting and replacing failed components.

C. Loss of NCAR earth segment availability due to rodent damage to outside earth station components.

Significant spacecraft time was lost to earth segment unavailability as the result of rodent damage to the power cabling for the Low Noise Amplifier (LNA) on the HDR terminal. The cabling connecting the outside LNA to the interior power supply was completely dissected by rodents. While it took a fair amount of time and effort to locate the source of earth station inoperability, repair was simple and complete functionality was restored quickly. Outside components were protected as well as practicable, however more intense hardening of facilities and routine, periodic inspection for such damage in future HDR implementations may help avoid such downtime.

D. Difficulty meeting project milestones as a result of organizational boundaries.

The magnitude of the experiment required a number of organizations with very different missions, goals, charters and management to interact. Government, state and private/commercial entities cooperated to operate and manage the significant resources involved in the experiment. While some bodies could dedicate sufficient numbers of personnel and resources to operate the various systems and objects, others were almost hamstrung to provide the minima. Differences in organizational procedures and policies had the capability to impede planning and execution of the complex and interdependent facets of the experiment. One entity's difficulty completing a task or procedure in a timely manner would have an adverse or domino effect on the others.

No overall project management for the experiment was possible due to the distributed nature of the system and diversity of the participants. Loss of coordination of the separate and disparate project management entities could result in duplication of effort or loss of synergy. While regular meetings and telephone conference calls were held, progress was sometimes irregular and strained. The inability to cross organizational boundaries to affect action or progress during a critical phase often left project managers feeling helpless, exacerbated by an nearly impossible schedule.

While the intrusion into autonomous organizational domains by singular project management is neither possible nor desirable, future deployments of supercomputing experiments of the COOP 3D nature may benefit from a parallel or peer distributed project management architecture. In this scenario each entity could designate project managers having a similar level of management authority in their respective organizations. Temporary but similar management and reporting structures could be established at each organization which would mirror overall project structure, practices and goals. Through this an automatic synergy would be realized; eliminating much duplicated effort and standardizing procedures by establishing a ubiquitous infrastructure common to critical participants. Managers could then focus their team's efforts to resolve technical or procedural problems crossing organizational boundaries with out frustration while adhering to the commonly drawn plan, timeline and goals.

Conclusion

In spite of the difficulties the goals of the experiment were met. Adequate performance was obtained from a terrestrial/satellite hybrid network architecture to support interactive data communications between high performance end systems here-to-fore requiring low latency, high bandwidth interconnections.

New ideas were developed and proven which revitalized existing technologies. The venerable technologies of geostationary satellites were enhanced to provide high capacity, fiber optic quality communications in a new arena. Mature and proven by its functionality on the Internet, the classical TCP/IP protocol suite was successfully adapted to operate over modern, high capacity physical layer technologies. Successful integration of dissimilar physical layer architectures was achieved, permitting interoperability between legacy systems and modern fiber-based communications technologies.

Inexpensive and reliable high bandwidth satellite communications systems like ACTS could increase the the utility of high performance computing. Such systems could facilitate real-time, collaborative efforts in scientific research between non-collocated researchers anywhere in the world at any time without reliance on terrestrial systems. Earth and atmospheric scientists could access supercomputing resources immediately from sites in Lesser Developed Countries or remote areas not served by modern high capacity data communications, enabling them to advance the state of scientific discovery sooner and at less expense.

Future deployment and availability of wide band satellite communications systems will be an enabling technology of the emerging Broadband Integrated Services Digital Network (B-ISDN). These systems will provide cost effective, high capacity communicaitons for the integrated voice, data and imaging applications which will make up the National Information Infrastructure (NII) and Global Information Infrastructure (GII). The presence of such systems will accelerate the deployment of these backbones in the United States and worldwide.